MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1138 July 1989
C.B.I.P Memo No. 40

Stimulus familiarity determines recognition strategy for novel 3D objects

Shimon Edelman Heinrich Bülthoff Daphna Weinshall

Abstract

Everyday objects are more readily recognized when seen from certain representative, or canonical, viewpoints than from other, random, viewpoints. We investigated the canonical views phenomenon for novel 3D objects. In particular, we looked for the effects of object complexity and familiarity on the variation of response times and error rates over different views of the object. Our main findings indicate that the response times for different views become more uniform with practice, even though the subjects in our experiments received no feedback as to the correctness of their responses. In addition, the orderly dependency of the response time on the distance to a "good" view, characteristic of the canonical views phenomenon, disappears with practice. One possible interpretation of our results is in terms of a tradeoff between memory needed for storing specific-view representations of objects and time spent in recognizing the objects.

This report describes research done at the Massachusetts Institute of Technology within the Artificial Intelligence Laboratory and the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences and Whitaker College. The Center's research is sponsored by grant N00014-88-K-0164 from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Alfred P. Sloan Foundation; and by National Science Foundation grant IRI-8719392. The Artificial Intelligence Laboratory's research is sponsored by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76-85-C-0010 and in part by ONR contract N00014-85-K-0124. Dr. Edelman and Dr. Weinshall are supported by Chaim Weizmann Postdoctoral Fellowships from the Weizmann Institute of Science and by a NSF Presidential Young Investigator Award to Professor Ellen C. Hildreth. Prof. Bülthoff is currently at the Department of Cognitive and Liguistic Sciences, Brown University.

1 Introduction

A common approach to the study of visual recognition postulates that there exist in the visual system representations of familiar objects and scenes. To recognize an object, the system compares it with each of the stored models. Such a comparison would appear possible only after the input image and the stored representations are brought to a common form. Consequently, the nature of the representation must be reflected in the performance of the system [1].

One possibility is that the visual system stores a few representative (canonical) views of each known object, along with the information that permits it to normalize the appearance of an input object by computing how it would look like from a canonical viewpoint [2]. Palmer, Rosch and Chase [3] found that canonical views of commonplace objects can be reliably characterized using several criteria. For example, when asked to form a mental image of an object, people usually imagine it as seen from a canonical perspective. In recognition, canonical views are identified more quickly than others, with response times decreasing monotonically with increasing subjective goodness [3].

This dependency of response time on the distance to a canonical view is expected if one draws an analogy between recognition by viewpoint normalization on one hand ([4], [5]) and mental rotation on the other hand ([6], [7]). The very existence of canonical views may then be attributed to a tradeoff between the amount of memory invested in storing object representations and the amount of time that must be spent in viewpoint normalization. Remembering a frequently encountered view of an object may lead to its faster recognition in subsequent encounters.

By the same argument, no preferred perspective should exist for familiar objects that are equally likely to be seen from any viewpoint. Indeed, there is evidence that normalization effects on recognition latency (as reflected in the existence of preferred views) disappear with practice for a variety of 2D stimuli such as line drawings of common objects [8], random polygons [9], pseudo-characters [10] and stick figures [11].

The aim of the present work is to explore further the issue of canonical views in object recognition. Our method differs in several respects from previous studies.

1. Our stimuli are images of novel 3D objects with controlled complexity. This facilitates the study of the effects of object complexity and familiarity on recognition.

2. The stimuli appear in various 3D orientations, bringing the experimental viewing conditions closer to those of real-world vision.

3. Our task involves recognition in the sense that it requires that the subject compare a displayed object with a target object previously committed to memory. In most earlier studies, subjects had to detect whether the displayed object was familiar or novel, or to make a handedness decision, such as whether the displayed object was a mirror image of the target.

4. Subjects are not required to name the stimuli. This reduces the number of different cognitive modules required for solving the task, making the reaction time correspond more closely to the actual duration of recognition.

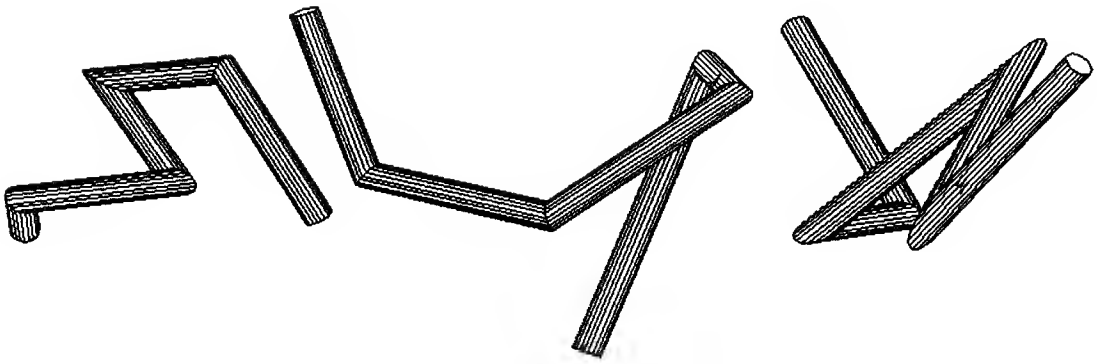

Figure 1: Examples of wire-like objects. Shaded, grey-scale images of similar wires were used as stimuli in the experiments.

2 Experimental Paradigm

Let us define the viewpoint coordinates of an observer with respect to an object, θ and ϕ, as the longitude and the latitude of the eye (or the camera) on an imaginary sphere centered at the object. One would expect a function $R(\theta, \phi)$ measuring the ease of recognition for a 3D object to possess one or more peaks, corresponding to its canonical views. We assessed the dependency of R on the object's complexity and on its familiarity to the subject, in a two-alternative forced-choice reaction time paradigm. Two measures of R, reaction time and error rate, were used.

2.1 Stimuli

We used the Symbolics S-Geometry^{TM} 3D graphics package to generate novel wire-like objects of small, nonzero thickness (Figure 1). This permitted us to simulate surface shading, while minimizing object self-occlusion. The objects were created in two steps. First, a straight five-segment chain of vertices was made. Second, each vertex was displaced in 3D by a random amount, distributed normally around zero. By definition, the variance of the displacements determined the complexity of the resulting wire. Third, the size of the resulting object was scaled, so that all the wires were of the same length.

Thirty novel 3D objects, generated according to the procedure described above and grouped by average complexity into three sets of ten, served as stimuli in the experiment (Figure 2). 144 evenly spaced images of each of the objects were produced by stepping the camera[1] by 30° increments in latitude and longitude. The images were rendered with the Symbolics S-Render^{TM} program, using the Lambertian surface reflectance model, with a point light source of intensity 1.0 (located at the camera) and an ambient light source of intensity 0.3. During the experiments, the images were displayed on a CRT monitor, on a dark background, under subdued ambient illumination. The images subtended an angle of approximately 6° at a distance of 120 *cm*.

[1] Here and below we refer to the simulated camera, light sources, etc.

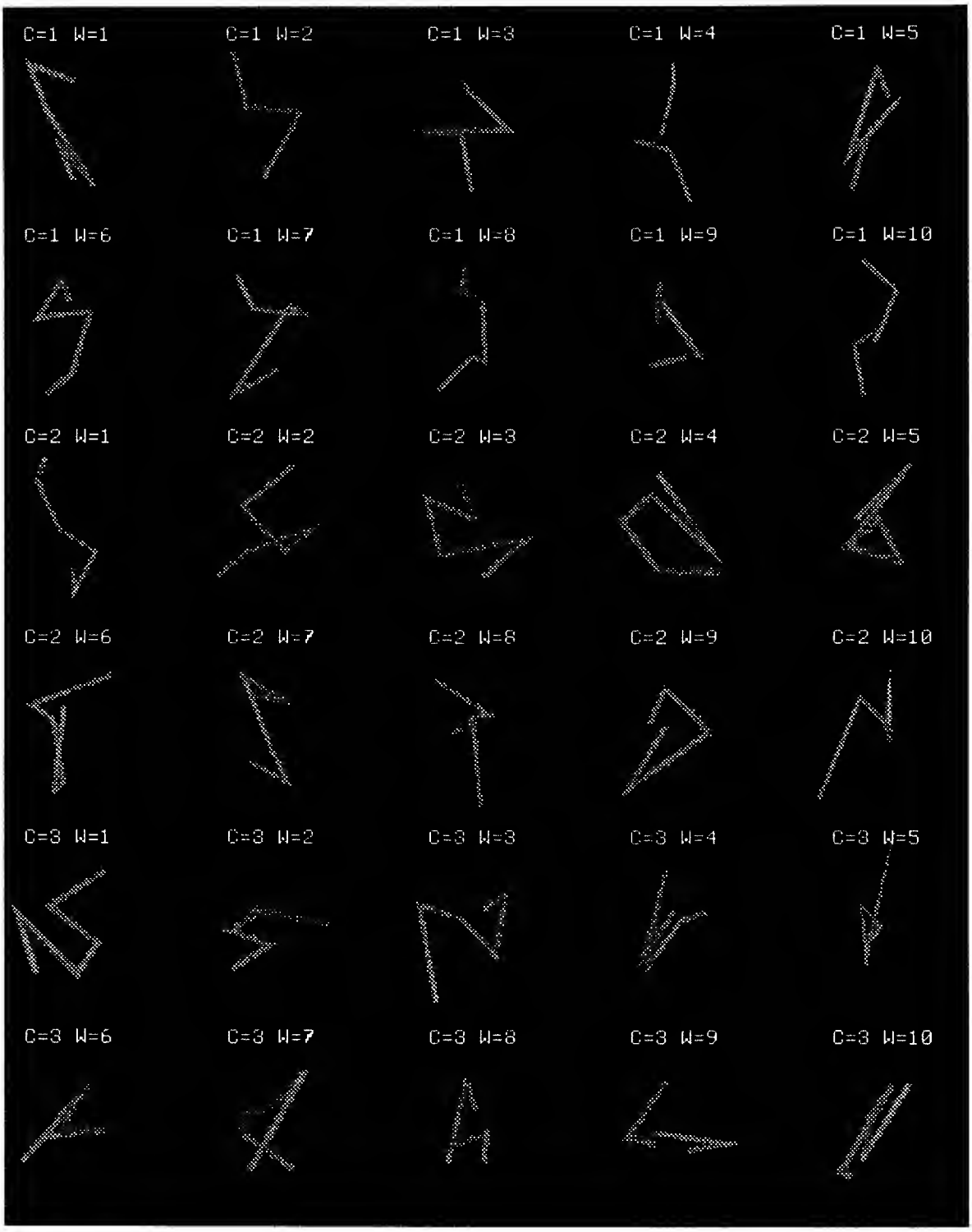

Figure 2: Thirty novel 3D wire-like objects were used in the experiment. The wires were grouped by complexity into three sets of ten. In this figure, the wires are marked by complexity (C=1, 2 or 3, corresponding to the low, middle and high complexities) and wire number (W, between 1 and 10).

2.2 A Pilot Experiment: Subjective Judgement

One of the operational definitions of a canonical view of an object, originally put forward by Palmer et al. [3], involves subjective judgement. When people are asked to rate the relative "goodness" of different views of everyday objects, the ratings tend to be highly correlated. In other words, there appears to exist a standard notion of what constitutes a good (informative, easy to recognize) view of familiar objects such as houses and horses. In our first experiment, we looked for a similar consensus in the domain of wire objects.

Four subjects rated 16 fixed views of each of the 10 test objects constituting the middle complexity set (see Figure 2) on a scale of 1 to 7 (the worst and the best ratings, respectively). The subjects were first allowed to familiarize themselves with the stimuli, by rotating them on the CRT display, using the computer keyboard. The subjects then interactively chose the best view for each object and were asked to rate the 16 standard views. The rest of this section describes the outcomes of three different analyses of the results. In addition, the subjective best-view information was used in the analysis of a subsequent experiment, along with objective best-view information, obtained from reaction time data.

The ratings for each view of each object were subjected to a 2-way nested effects (View(Object) $\times$ Subject) analysis of variance (ANOVA). For most objects, the effects of View and of Subject were highly significant. Two exceptions were object #6, all of whose views received similar ratings (View: $F(15,63) = 1.36$, $p > 0.21$; Subject: $F(3,63) = 2.25$, $p < 0.095$) and object #9, about whose mean rating there was the highest consensus among subjects (View: $F(15,63) = 2.53$, $p < 0.0082$; Subject: $F(3,63) = 1.41$, $p > 0.25$).

A different way to assess the agreement among subjects is to compute the correlations between the 16-tuples of standard-view ratings. Averaged over all 10 objects, the correlation was quite high (Kendall coefficient of concordance 0.45), although much lower than the figures reported by Palmer et al. [3]. The outstanding objects were once again #6 and #9, for which some of the pairwise inter-subject correlations were as low as -0.39. Comparing this with the ANOVA results, we note that of all objects those that had the most uniform mean ratings also yielded by-view ratings that were least correlated among subjects. In other words, when asked to rate the views of an object that looked much the same from every viewpoint, subjects tended to come up with quite noisy ratings.

We have employed principal factor analysis to look for possible patterns in the assignment of subjective goodness ratings. For 8 out of 10 objects, the FACTOR procedure retained just two principal factors. For objects 2 and 4, three factors were retained, although the variance due to the third factor was in each case much smaller than the variance due to the second one. The outcome of this analysis suggests that the number of different criteria used by the subjects in the assignment of goodness ratings is as small as two.

3 Recognition Experiments

Canonical views of an object may also be defined as those views that yield relatively short response latencies in a recognition task [3]. If, as previous experimental evidence suggests ([8], [9], [10], [11]), the advantage of some views over the others is linked to the subject's familiarity with the stimulus, one should expect the strength of the canonical views phenomenon to depend on familiarity, e.g., as depicted in Figure 3.

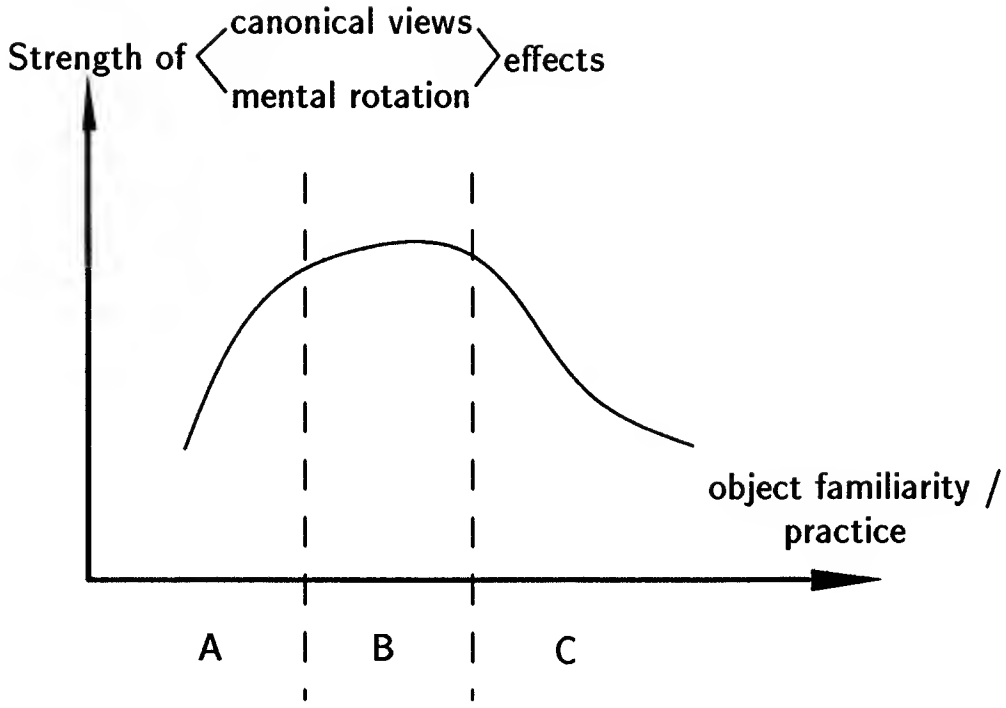

Figure 3: Expected influence of object familiarity on the canonical views phenomenon and on the strenth of response latency effects related to mental rotation (see text). Both unfamiliar and highly overlearned objects should yield relatively uniform reaction times when recognized from different viewpoints (regions A and C). For objects that are frequently seen from some, but not from all, viewpoints (region B) there should be a relatively strong dependence of response latency on viewpoint.

A measure of the strength of the canonical views phenomenon can be obtained by dividing the standard variation of response latency over different views of an object by the mean latency for that object[2]. What mechanism could bring about a decrease in this measure? A basic prerequisite seems to be the capability of the subjects' visual system to be imprinted with views of arbitrary objects[3].

In the experiment we describe next, we looked for, and found, evidence that could signify imprinting with novel views. We exposed the subjects repeatedly to the same small set of views of the stimuli, leaving the question of the transfer of recognition from familiar to novel views for future research. Consequently, we expected the variation of response latency over views to decrease with practice.

3.1 Method

Thirty novel wire objects (see Figure 2) served as stimuli. The basic experimental run used ten objects of the same complexity and consisted of ten blocks, in each of which a different object was defined as the target for recognition. Each block had two phases:

Training: In the beginning of each block, the subject was shown all 144 views of the target twice, in a natural succession. The target was perceived as tumbling in space, with the kinetic depth effect contributing to the three-dimensional appearance of the object.

Testing: In the rest of the block, a subset of 16 fixed views (spaced by 90° in latitude and longitude) was used for each object. The subject was presented with a sequence of stimuli, shown one at a time. Half of these were views of the target. The other half were views of the rest of the objects from the current set.

The appearance of a stimulus was preceded by a fixation point. The stimuli stayed on until the subject responded. The response times were measured in a two-alternative forced choice paradigm. The subject had to press one key if the displayed object was the current target, and another key otherwise. No feedback was given as to the correctness of the response.

3.2 Experiment 1

Three experienced subjects (the authors) participated in the first experiment. The basic run has been repeated three times (once per complexity group, in a fixed order) over a period of a few

[2]Below, we employ an additional, different, measure that has a bearing also on the phenomenon of mental rotation.

[3]A related question is, can people recognize an object from a novel, radically different, viewpoint [12]? Theoretically, the structure from motion theorems ([13], [14]) indicate that it should be possible to reconstruct the 3D shape of an object, and hence its appearance from an arbitrary viewpoint, given enough discrete views. In practice, there are indications that the visual system's ability for such reconstruction is limited ([15], [12], [16], [17]). A positive answer to this as yet unresolved empirical question would strengthen the prediction of eventual uniformity of response latency over views. A negative answer, on the other hand, would mean that the visual system is more like an associative memory than a general-purpose computational device. In that case, the response latency to novel views would remain high. This could contribute to high overall variation of latency over views, but only if novel, as well as familiar, views are tested. Note that in any case a decrease in the variation over familiar views is expected. An initially high variation of latency that decreases relatively slowly with practice would lend more support to the associative memory interpretation.

days. Altogether, 14400 responses were obtained. Each of the 16 views of every target appeared during the test phase five times. We refer to the first three and the last two appearances of each view, respectively, as "session 1" and "session 2".

In the following analysis we used only the data from those observations in which the stimulus shown was actually the target (as opposed to one of the non-targets). This considerably simplified the analysis, at the expense of wasting some data[4]. Latencies of correct responses (response times or RTs) and error rates (ERs) were averaged to yield a single value per session per view per object. RTs longer than 3 *sec* or shorter than 250 *msec* were discarded. Mean ER was 13.2%.

To find out whether there was any significant time/accuracy tradeoff, we have correlated RT and ER data, averaged over views of the objects. A time/accuracy tradeoff would be expressed in a negative correlation coefficient. For two of the subjects (SYE and DW) significant positive correlation was found between the RT and the ER data (SYE: $r = 0.66$, $p < 0.0001$; DW: $r = 0.53$, $p < 0.0027$). For the third subject, the correlation was close to 0 (HHB: $r = 0.06$, $p > 0.7$). Between-subjects correlations were high, except for the RTs of the third subject. Thus, for two subjects evidence against time/accuracy tradeoff was found, while for the third subject there appeared to be no connection between response times and error rates.

The decrease of the mean RT with practice was a basic effect that we had expected to find. This effect would have masked any differential effects of familiarity on the recognition of objects from different viewpoints, unless a measure of canonicality (the advantage of some views over others) insensitive to the overall decrease in mean RT were used. We chose the *coefficient of variation* of RT over the different views (defined as the ratio of the standard deviation of RT to the mean of RT) as one measure of the strength of the canonicality effect, and used analysis of variance to find its dependency on familiarity.

A different way to assess the canonical views effect is by looking for an explicit dependency of the RT on the attitude of the object relative to the observer. In this case data cannot be pooled over different objects, unless a common reference attitude is defined. One possibility is to define the (subject-specific) best view for each object as the view with the shortest RT. One could then characterize RT as a function of object attitude by measuring its dependency on $D = D(subject, target, view)$, the distance between the best view and the actually shown view[5] We used regression analysis to characterize $RT(D)$ and $ER(D)$.

3.2.1 Analysis of response times and error rates

Figure 4 shows plots of RT and ER vs. Target, grouped by Subject and Complexity. We include these plots for completeness only, since, as we have argued above, it is the variation, rather than the mean, of RT and ER over different views of an object that is especially relevant to the issue of canonical views.

[4]The target object was shown on one half of the trials (the target trials). On the non-target trials, the other nine objects appeared with an equal likelihood. Were the data for those trials included in the analysis, the data set would become unbalanced.

It would be interesting to analyze the data from the non-target trials separately, mainly to look for a pattern in the errors (of the false alarm type) made on those trials. The results of this analysis could be presented as a confusion table, a common format in the study of, e.g., letter recognition.

[5]We define D between two views, $v_1 = (\theta_1, \phi_1)$ and $v_2 = (\theta_2, \phi_2)$, as the city-block distance in the θ, ϕ (longitude, latitude) coordinates of the viewing sphere: $D(v_1, v_2) = \max\{|\theta_1 - \theta_2|, |\phi_1 - \phi_2|\}$.

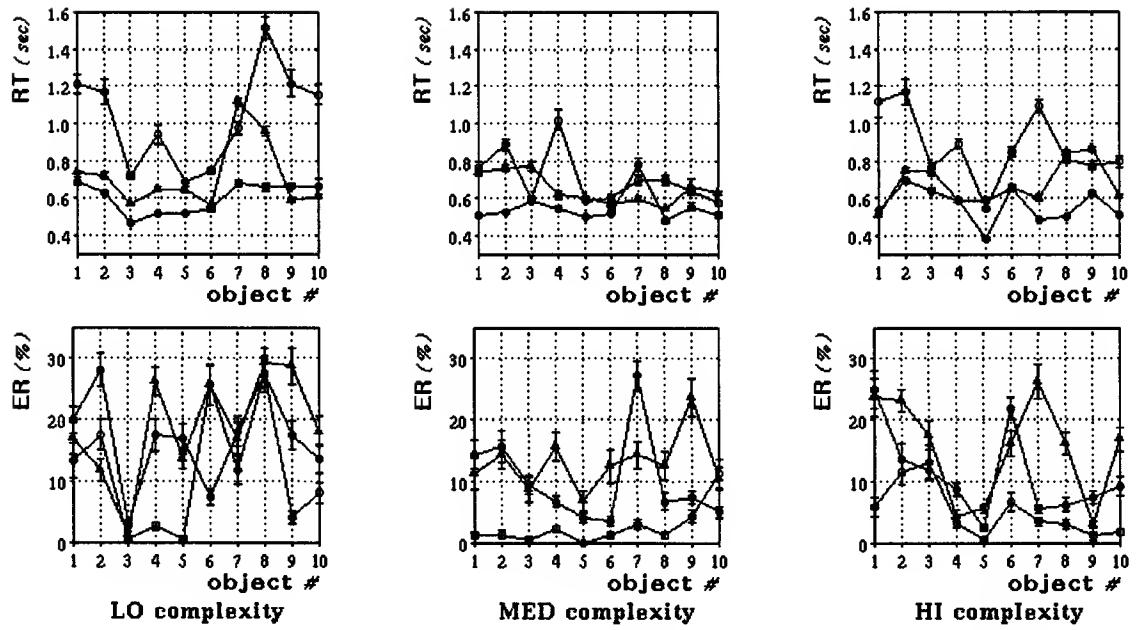

Figure 4: Experiment 1: response times (RT, *sec*) and error rates (ER, %) vs. Target, by Subject and Complexity. Curves from subjects DW, HHB and SYE are marked, respectively, by small circles, triangles and dots. The error bars denote standard deviation, computed over views for each object. In many cases, the bars are masked by the data point marks.

Mean RTs were 0.75, 0.69 and 0.62 *sec* for low, middle and high complexities. Grouped by session, the RTs were 0.71 and 0.66 *sec* for sessions 1 and 2, respectively[6]. The only significant interaction was that of Complexity × Subject. The ranking of the subjects by RT was (from the highest to the lowest) DW, HHB, SYE.

The mean ER for low complexity was 17.9%, for high complexity – 12.0%, and for middle – 9.7% (the last difference was not significant). The mean ER in session 2, 15.2%, was higher than in session 1, 11.2%. The ranking of subjects by ER was HHB, SYE, DW. Note that it is different from the ranking by RT.

Although considerable variation across subjects is apparent in Figure 4 and the next two plots, analysis of the normalized variation of RT and ER over views (described below) revealed no interaction between Subject and the other independent variables of interest, Complexity and Session. In other words, although the means of RT and ER by Subject varied, the effects of Complexity and Session on the strength of the canonical views phenomenon were similar for all subjects.

3.2.2 Analysis of the coefficient of variation of response time and error rate

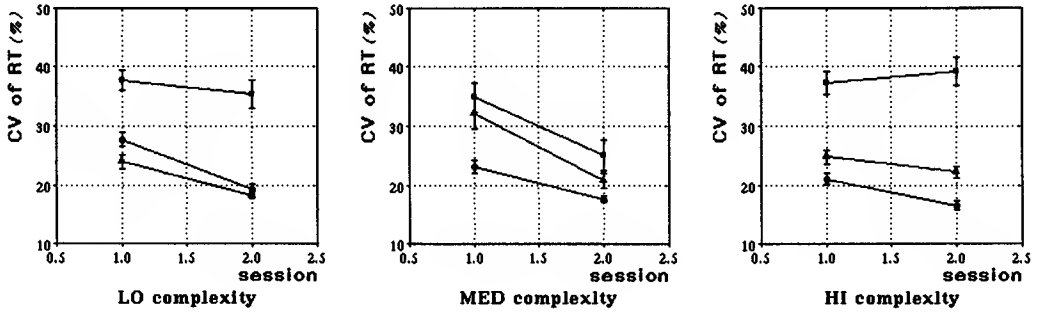

Figure 5: Experiment 1: coefficient of variation of RT (%) over views for the two sessions, by Subject and Complexity (square, triangle and dot stand for DW, HHB and SYE). The decrease of the c.v. of RT with Session is significant.

The coefficient of variation of RT over different views of objects decreased with practice (see Figure 5). Effects of Subject and Session, but not of Complexity, were significant. All three means by Complexity were close to 26%. The means by Session were 29.1% and 23.8% for sessions 1 and 2.

For ER (see Figure 6), all main effects were significant. The means of the coefficient of variation of ER by Complexity were 156%, 186% and 206% for low, high and middle sets, respectively (the last difference was not significant). The means by Session were 168% and 198% for sessions 1 and 2.

3.2.3 Regression analysis of RT, ER

[6] All differences among the means reported here and below were found significant by Duncan's multiple-range test at $p < 0.05$, unless otherwise noted. See the appendix for more detailed results.

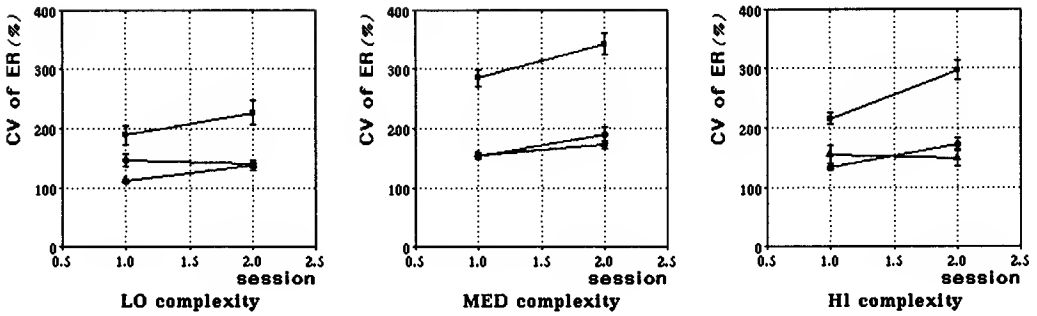

Figure 6: Experiment 1: coefficient of variation of ER (%) over views for the two sessions, by Subject and Complexity (square, triangle and dot stand for DW, HHB and SYE). The effect of Session is significant, mainly due to DW's contribution.

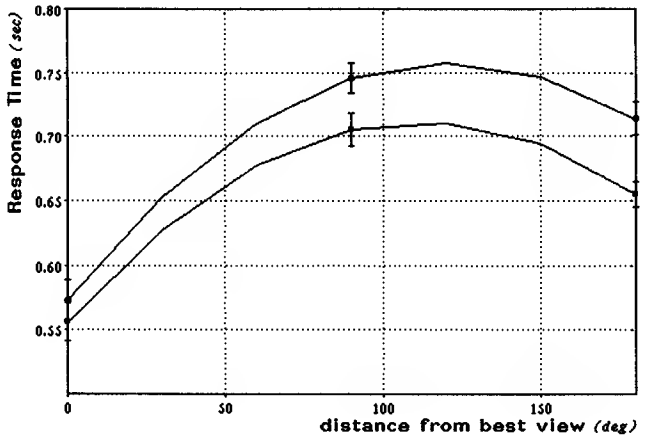

Figure 7: Regression curves of RT on D for the two sessions of experiment 1. Data are pooled over subjects. Means and standard errors of over 1000 points are shown. RT is measured in *sec*, D – in multiples of 30^o. $D = 0$ corresponds to the best view. The lower curve refers to session 2. Error bars denote twice the standard error of the mean for the corresponding points.

Regression analysis yielded a significant quadratic component. For session 1, the dependency of RT on the distance D between the displayed view and the best view of the object (as defined by the subjects' performance in this experiment, not by the subjective judgement in experiment 1; see below) was $RT = 0.576 + 0.095D - 0.013D^2$, where RT is measured in seconds and D – in increments of 30°. The dependency remained significant for session 2: $RT = 0.558 + 0.076D - 0.010D^2$. Notably, the regression of ER on D and D^2 was not significant, either for session 1, or for session 2.

The outcome of the regression analysis is not trivial, in the sense that, in principle, the RT can vary with view in a disorderly fashion. In that case, no consistent variation of RT with the distance to the best view would be revealed by the analysis. Indeed, the regression of RT on the distance to a *random* view (fixed for each object and subject), computed as a control, was not significant. Neither was the regression of RT on the distance to the subjectively best view, as obtained in the subjective judgement experiment (this probably indicates that the subjects' intuition as to what constitutes a "good" view of a wire object is poor, at least in comparison to Palmer's results for common objects [3]).

The shapes of the regression curves of RT for the two sessions of experiment 1 seem to be different (see Figure 7). A multivariate test of the difference between the two sets of regression coefficients[7] came short, however, of confirming this impression. This was the main reason for carrying out experiments 2 and 3.

3.3 Experiment 2

In this experiment, one of the original subjects (SYE) was tested repeatedly, to elucidate the dependency of regression results on object familiarity. For this subject, the responses of both sessions of the previous experiment, consisting together of 5 trials per view per object, were combined, and an additional 5-trial session was performed. The results of this experiment appear below.

3.3.1 Analysis of the coefficient of variation of RT and ER

The plot of the coefficient of variation of RT for experiment 2 (Figure 8) shows that it decreased with session for the low and the medium, but not for the high, complexity groups. The overall effect of session was weak, but noticeable (experiment 3, described below, confirmed this effect).

The plot of the coefficient of variation of ER for experiment 2 appears in Figure 9. Only the main effect of complexity was significant. A separate analysis for session by complexity revealed no significant effects of session in any complexity group.

3.3.2 Regression analysis of RT, ER

Regression of RT on D and D^2 for session 1 (see Figure 10) was significant, giving $RT = 0.475 + 0.058D - 0.007D^2$. Importantly, it was not significant for session 2. That is, the dependence of RT on the distance to the best view was strongly diminished. Regression of ER was not significant for both sessions.

[7] Excluding the intercepts – we were not interested in mere uniform decrease of RT for all views.

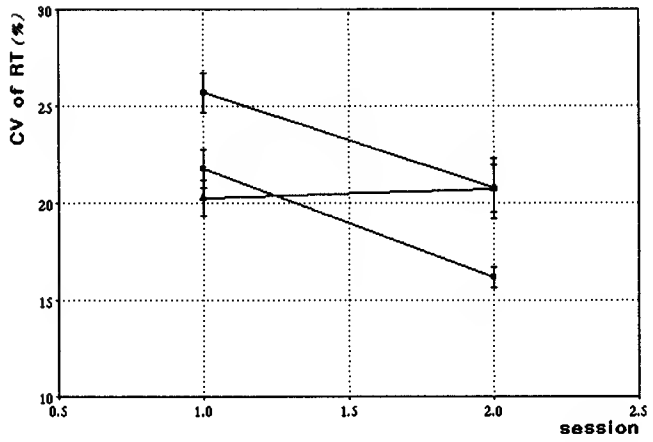

Figure 8: Coefficient of variation of RT over views (%) for the two sessions of experiment 2, by complexity (dot, square and triangle mark low, middle and high complexity, respectively).

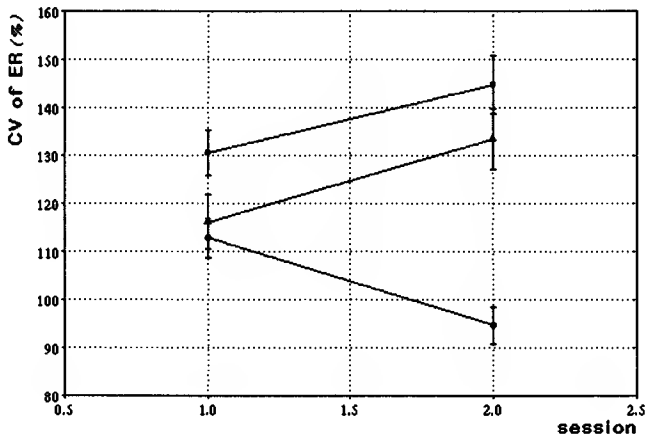

Figure 9: Coefficient of variation of ER rate over views (%) for the two sessions of experiment 2, by complexity (dot, square and triangle mark low, middle and high complexity, respectively).

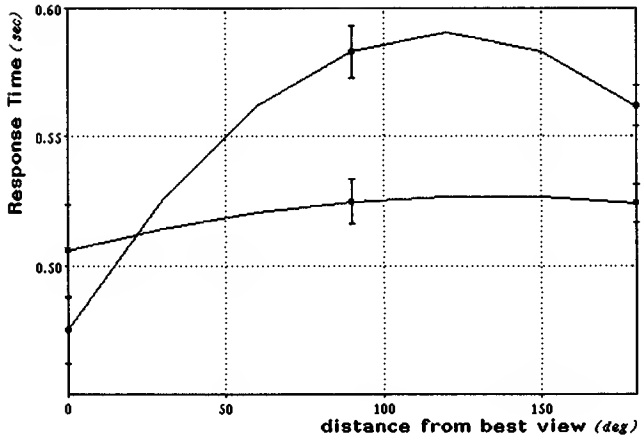

Figure 10: Regression curves of RT on D for the two sessions of experiment 2. Scale labeling is as in the previous regression plot. The flatter curve refers to session 2. Error bars denote twice the standard error of the mean for the corresponding points.

3.4 Experiment 3

One lesson from the previous two experiments is that at least 10 exposures per view per object are necessary to obtain a clear effect of object familiarity on the strength of the canonical views phenomenon. Having demonstrated this effect with two 5-trial sessions in an experiment with one subject, we repeated the experiment with four additional (naive) subjects, to improve the statistical significance of the results. Thus, after experiment 3 we had data for five subjects, each of whom was tested on ten different objects (the middle complexity set), in two 5-trial sessions.

The dependency of the coefficient of variation of RT on session in this experiment is illustrated in Figure 11. Mean c.v. of RT decreased from 36.6% in session 1 to 26.6% in session 2. This effect was highly significant. The effect of Subject was also significant (the means of c.v. of RT by Subject ranged from 19% to 40%), but, importantly, there was no Subject $\times$ Session interaction.

The plot of the coefficient of variation of ER vs. session in experiment 3 appears in Figure 12. The means of c.v. of ER are 140% and 126% for sessions 1 and 2, respectively. Except for one subject, there is no significant effect of Session. The effect of Subject here is significant, and so is the Subject $\times$ Session interaction. The overall effect of Session is not significant. In general, these results are close to those of the previous experiments.

As in previous experiments, the dependency of RT on D and D^2 for session 1 (see Figure 13) was obvious for session 1, giving $RT = 0.604 + 0.079D - 0.009D^2$, but negligible for session 2. Regression of ER was not significant for both sessions.

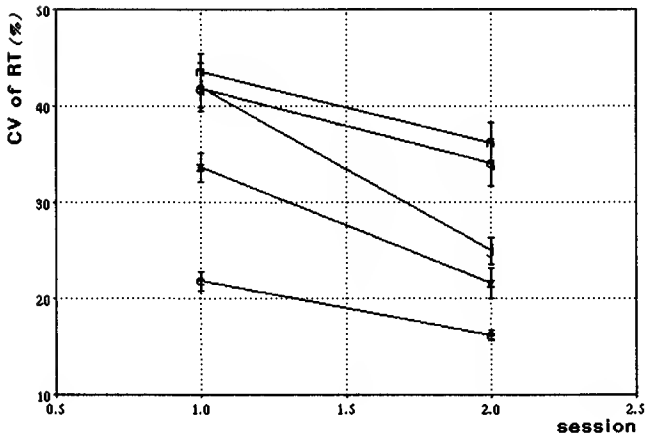

Figure 11: Coefficient of variation of RT over views (%) for the two sessions of experiment 3 by Subject.

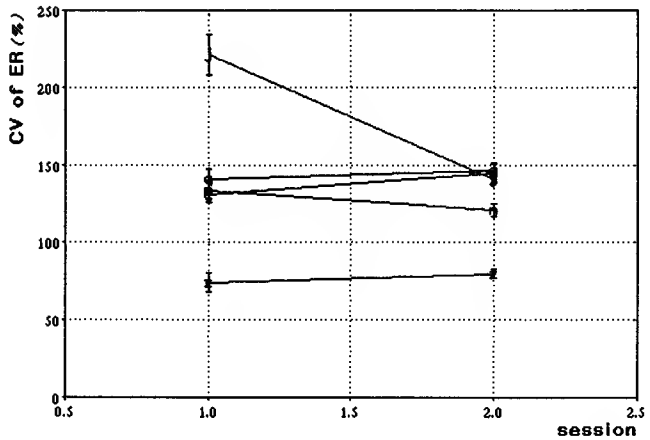

Figure 12: Coefficient of variation of ER rate over views (%) for the two sessions of experiment 3, by subject.

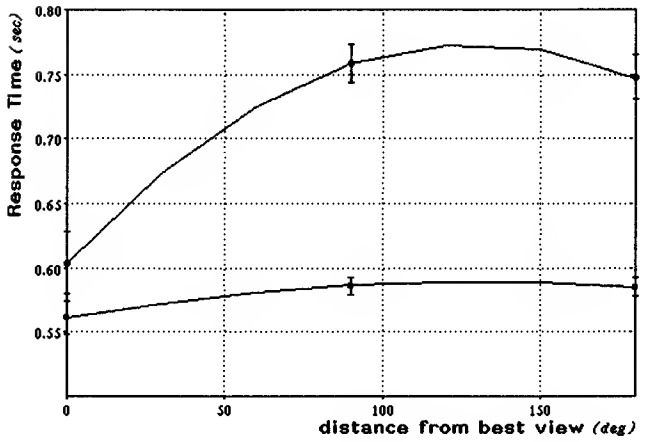

Figure 13: Regression curves of RT on D for the two sessions of experiment 3. Scale labeling is as in the previous regression plot. The flatter curve refers to session 2. Error bars denote twice the standard error of the mean for the corresponding points.

4 Discussion

4.1 Complexity effects

The influence of stimulus complexity on mean RT and ER was in part expected (higher complexity resulted in longer RT and higher ER than middle complexity), and in part unexpected (lower complexity had a similar effect). A possible explanation involves the notion of viewpoint-invariant, non-accidental features of 3D objects, e.g. parallel lines, collinear points and co-terminating segments [4]. In our case, these features are more likely to be present in wire objects that have higher complexity (see the description in section 2.1 of the procedure we used to generate the stimuli). While the presence of features such as collinear segments can facilitate recognition, having too many of them would have an opposite effect, e.g., by prompting the subject to resort to a more complicated procedure. Having too few of these features could also impede recognition (by increasing ambiguity).

Stimulus complexity had no effect on the coefficient of variation of RT over views. It appears that most of the variation of RT (as opposed to the mean RT) is due to factors other than complexity, such as the general outlook of our stimuli (e.g., an elongated wire seen end-on would be naturally harder to recognize than the same wire seen from the side). On the other hand, stimulus complexity affected the coefficient of variation of ER over views. We do not attempt to interpret this effect, because of the possible Subject $\times$ Complexity interaction (see the difference between the data for subject DW and the other two subjects in Figure 6).

4.2 Session (familiarity) effects

Our data indicate a clear effect of familiarity on the prominence of canonical views, at least for the kind of objects we have used as stimuli. Familiarity appears to reduce the differences in

RT among different views of the object (see Figures 5, 8 and 11), and to render insignificant possible effects of mental rotation, as manifested in the dependency of RT on the distance to the canonical view (Figures 10 and 13). The variation of ER over views does not seem to change with practice. Of the seven subjects we have tested, the data from one exhibited an increase in the variation of ER with practice, one subject showed a decrease, and for the other five subjects the variation of ER did not change (Figures 6 and 12).

We interpret session effects on RT in the absence of feedback as an indication of *imprinting* of familiar views that happens merely as a result of repeated exposure. As a result of the imprinting, the response times for different views of the same object become more uniform, whereas the variation in the error rates appears not to be affected.

4.3 Interpreting regression results

Experimental results in which recognition time of an object depended on the amount of rotation necessary to bring it to a familiar orientation have been previously interpreted in terms of mental rotation [16]. The major argument in favor of this interpretation is indirect and has to do with similarity between the slope of the regression curve in recognition and in classical mental rotation tasks ([18], [7]). The reciprocal of the coefficient of D in the regression equation for RT(D) in session 1 in our experiments (approximately 300 deg/sec) is also consistent with that of mental rotation.

This result, along with the apparent absence of an orderly dependence of ER on D, can be accommodated by a theory of recognition that involves two distinct stages: normalization and comparison (cf. Ullman's recognition by alignment [5]). In the normalization stage, the image and a model are brought to a common attitude in a visual buffer. This operation could be done by a process analogous to mental rotation, which would take time proportional to the attitude difference between the image and the model. Subsequently, a comparison would be made between the two. The time to perform the comparison could depend, e.g., on the object's complexity, but not on its attitude, so that the comparison stage would contribute a constant amount to the overall recognition time. On the other hand, the error rate of recognition would be largely determined by the comparison stage. With practice, more views of the stimuli could be retained by the visual system, resulting in a smaller average amount of rotation necessary to normalize the input to a standard, or canonical, appearance. The response times for the initially "bad" views (determined by the normalization process) would decrease, reducing the variation of RT over views. On the other hand, the mean error rates for the "bad" views (determined by the comparison process), and, consequently, the variation of ER over views, would not change, because of the absence of feedback to the subject. This is compatible with our observations.

The strong quadratic component in the regression equations for RT(D) may signify the presence of more than one preferred, or canonical, view. Imagine the viewing sphere (see section 2) centered around a wire-like object, with the best (shortest-RT) view at the north pole. Then the view from the south pole of the sphere (which is at $D = 6$, or 180^o, from the north pole) ought to yield shorter RT than views from the equator, because the projection of a wire looks almost the same from two diametrically opposite directions. This may explain the shape of the regression curve for RT(D).

5 Summary

To recapitulate, our main findings are as follows.

- Stimulus complexity has no effect on the variation of RT over views;
- Stimulus familiarity reduces the variation of RT over views;
- Familiarity reduces the effect that can be interpreted in terms of mental rotation, namely, the dependency of RT on the distance to the canonical view.

These effects support the notion of a tradeoff between time required for viewpoint normalization and memory invested in storing multiple views of objects. Our subjects appear to possess an impressive capacity for remembering random views of novel objects. We believe that novel objects are most effectively remembered when they are important behaviorally (e.g., when they appear as targets in a recognition experiment). The nature of object representation in long-term memory has been the subject of a long debate in cognitive psychology (e.g. [19], [20]). The present paper described an investigation of one aspect of the representation problem — the effect of object familiarity on the canonical views phenomenon [3]. Several additional issues that we are currently exploring are (1) the amount of 3D information retained in specific-view representation; (2) the ability of the visual system to infer the appearance of objects from unfamiliar attitudes (cf. [15], [12]); (3) the visual vocabulary used to build object representations and (4) computational aspects of object representation.

One computational model of recognition that is consistent with our findings is the two-stage recognition by alignment [5]. A possible explanation of the familiarity effect in terms of alignment involves mental rotation of object representations that becomes unnecessary when many specific views of objects are stored as a result of practice. In a related work ([21], [22]) we show that a self-organizing model that has no built-in provisions for rotating arbitrary objects may suffice to account for our experimental results.

Acknowledgements

We thank Zili Liu for his assistance in running experiment 3, Jeremy Wolfe and Ellen Hildreth for their comments on a draft of this paper, and Tomaso Poggio and Shimon Ullman for useful discussions.

Appendix: ANOVA results for experiments 1, 2 and 3

Experiment 1

A three-way ANOVA of RT (Complexity $\times$ Subject $\times$ Session) revealed significant main effects (Complexity: $F(2,162) = 14.97$, $p < 0.0001$; Subject: $F(2,162) = 64.0$, $p < 0.0001$; Session: $F(1,162) = 5.51$, $p < 0.02$). For ER, only the main effects were significant (Complexity: $F(2,162) = 14.18$, $p < 0.0001$; Subject: $F(2,162) = 21.57$, $p < 0.0001$; Session: $F(1,162) = 9.24$, $p < 0.003$). The means of RT and ER appear in Tables 1 and 2.

A three-way ANOVA of the coefficient of variation of RT (Complexity $\times$ Subject $\times$ Session) showed significant main effects for Subject ($F(2,162) = 30.74$, $p < 0.0001$) and Session

	DW	HHB	SYE
High	0.84	0.66	0.55
Med	0.97	0.71	0.58
Low	0.69	0.63	0.53

session	1	2
High	0.71	0.67
Med	0.64	0.60
Low	0.78	0.72

session	1	2
DW	0.87	0.80
HHB	0.68	0.65
SYE	0.58	0.53

Table 1: Mean reaction times (RTs, *sec*) in experiment 1 (from left to right: by Subject and Complexity; by Session and Complexity; by Session and Subject). See text for ANOVA results on the significance of the differences between the various means.

	DW	HHB	SYE
High	6.2	17.3	12.5
Med	3.1	14.4	11.5
Low	13.3	21.7	18.7

session	1	2
High	9.8	14.2
Med	8.3	11.1
Low	15.6	20.2

session	1	2
DW	6.4	8.7
HHB	15.0	20.1
SYE	12.3	16.2

Table 2: Mean error rates (ERs, %) in experiment 1 (from left to right: by Subject and Complexity; by Session and Complexity; by Session and Subject). See text for ANOVA results on the significance of the differences between the various means.

($F(1,162) = 12.06$, $p < 0.0007$), but not for Complexity. For the coefficient of variation of ER, all three main effects were significant (Complexity: $F(2,150) = 7.65$, $p < 0.0007$; Subject: $F(2,150) = 38.68$, $p < 0.0001$; Session: $F(1,150) = 7.19$, $p < 0.008$; the smaller number of degrees of freedom is due to missing values). Two interactions were noticeable (although not significant): Subject $\times$ Complexity ($F(4,150) = 1.55$, $p = 0.19$) and Subject $\times$ Session ($F(2,150) = 1.60$, $p = 0.20$). The means of the coefficients of variation of RT and ER appear in Tables 3 and 4.

	DW	HHB	SYE
High	38.3	23.5	18.8
Med	30.0	26.5	20.3
Low	36.6	21.1	23.5

session	1	2
High	27.7	26.0
Med	30.1	21.1
Low	29.8	24.3

session	1	2
DW	36.7	33.2
HHB	27.0	20.4
SYE	23.9	17.8

Table 3: Mean coefficient of variation (%) of reaction times over views in experiment 1 (from left to right: by Subject and Complexity; by Session and Complexity; by Session and Subject). See text for ANOVA results on the significance of the differences between the various means.

Regression of RT on D and D^2 for session 1 was significant ($F(2,1128) = 10.95$, $p < 0.0001$). It remained significant ($F(2,1128) = 9.16$, $p < 0.0001$) for session 2. The regression of RT on the distance to a *random* view (fixed for each object and subject), computed as a control, was not significant. The regression of ER on D and D^2 was also not significant, for either session. A multivariate test of the difference between the set of regression coefficients for session 1 and

	DW	HHB	SYE
High	261	152	152
Med	312	163	170
Low	206	125	143

session	1	2
High	165	207
Med	191	222
Low	149	164

session	1	2
DW	228	287
HHB	141	152
SYE	144	166

Table 4: Mean coefficient of variation (%) of error rates over views in experiment 1 (from left to right: by Subject and Complexity; by Session and Complexity; by Session and Subject). See text for ANOVA results on the significance of the differences between the various means.

that of session 2 (excluding the intercepts) was not significant ($F(2,1128) = 0.5$, $p = 0.6$).

Experiment 2

For RT, a one-way ANOVA for the effect of Session by Complexity gave $F(1,18) = 1.78$, $p < 0.2$ for low complexity; $F(1,18) = 6.47$, $p < 0.02$ for middle complexity; $F < 1$ for high complexity. The overall effect of Session in a two-way ANOVA (Complexity $\times$ Session) was weak, but present ($F(1,54) = 3.65$, $p < 0.06$).

For ER, a two-way ANOVA (Complexity $\times$ Session) showed only the main effect of Complexity as significant ($F(2,54) = 5.46$, $p < 0.007$). A one-way ANOVA for Session by Complexity revealed no significant effects of Session in any complexity group.

Regression of RT on D and D^2 was significant ($F(2,430) = 7.3$, $p < 0.0007$) for session 1, but not for session 2 ($F < 1$). Regression of ER was not significant for both sessions.

Experiment 3

For the coefficient of variation of RT, a two-way ANOVA, Subject $\times$ Session, showed significant main effects for Subject ($F(4,98) = 12.0, p < 0.0001$) and Session ($F(1,98) = 20.5, p < 0.0001$). The interaction was not significant ($F < 1$).

For the coefficient of variation of ER, the effect of Subject was strong ($F(4,98) = 16.9$, $p < 0.0001$) and of Session – marginal ($F(1,98) = 2.8$), $p = 0.1$; all of this due to one subject's contribution). The interaction was significant ($F(4,98) = 4.6$, $p < 0.002$).

Regression of RT on D and D^2 in session 1 was noticeable, but weak, due to considerable variability among subjects: $RT = 0.604 + 0.079D - 0.009D^2$ ($F(2,729) = 5.1$, $p < 0.0063$). In session 2, the regression was insignificant ($F < 1$).

References

[1] D. Marr. *Vision.* W.H. Freeman, San Francisco, CA, 1982.

[2] S. E. Palmer. The psychology of perceptual organization: a transformational approach. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and machine vision*, pages 269–340. Academic Press, New York, 1983.

session	1	2
SYE	21.8	16.2
JIN	42.0	24.9
NL	43.6	36.2
QI	41.9	34.1
ZH	33.7	21.6

session	1	2
SYE	130.5	144.7
JIN	221.1	140.5
NL	133.5	121.0
QI	140.4	146.1
ZH	74.4	79.9

Table 5: Mean coefficient of variation (%) of reaction times (left) and error rates (right) over views in experiment 3. See text for ANOVA results on the significance of the differences between the various means.

[3] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In J. Long and A. Baddeley, editors, *Attention and Performance IX*, pages 135–151. Erlbaum, Hillsdale, NJ, 1981.

[4] D. G. Lowe. *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston, MA, 1986.

[5] Shimon Ullman. An approach to object recognition: Aligning pictorial descriptions. A.I. Memo No. 931, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1986.

[6] R.N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[7] R. N. Shepard and L.A. Cooper. *Mental images and their transformations.* MIT Press, Cambridge, MA, 1982.

[8] P. Jolicoeur. The time to name disoriented objects. *Memory and Cognition*, 13:289–303, 1985.

[9] A. Larsen. Pattern matching: effects of size ratio, angular difference in orientation and familiarity. *Perception and Psychophysics*, 38:63–68, 1985.

[10] A. Koriat and J. Norman. Mental rotation and visual familiarity. *Perception and Psychophysics*, 37:429–439, 1985.

[11] M. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21, 1989.

[12] I. Rock, D. Wheeler, and L. Tudor. Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21:185–210, 1989.

[13] Shimon Ullman. *The interpretation of visual motion.* MIT Press, Cambridge, MA, 1979.

[14] R.Y. Tsai and T.S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. Technical Report R-921, Univ. of Illinois, Urbana-Champaign, 1981.

[15] I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293, 1987.

[16] M. J. Tarr. *Orientation dependence in three-dimensional object recognition.* PhD thesis, Dept. of Brain and Cognitive Sciences, MIT, 1989.

[17] S. Edelman, 1989. unpublished observations.

[18] S. Shepard and D. Metzler. Mental rotation: effects of dimensionality of objects and type of task. *J. Exp. Psychol.: Human Perception and Performance*, 14:3–11, 1988.

[19] S.M. Kosslyn. *Image and mind.* Harvard Univ. Press, Cambridge, MA, 1980.

[20] Z. Pylyshyn. *Computation and cognition.* MIT Press, Cambridge, MA, 1985.

[21] S. Edelman, D. Weinshall, H. Bülthoff, and T. Poggio. A model of the acquisition of object representations in human 3d visual recognition. In *Proc. NATO Advanced Research Workshop on Robots and Biological Systems*, Lucca, Italy, 1989. Springer Verlag. to appear.

[22] D. Weinshall, S. Edelman, and H. Bülthoff. A self-organizing multiple-view representation of 3d objects. A.I. Memo No. 1146, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989. in preparation.

CS-TR Scanning Project
Document Control Form

Date : 1/19/95

Report # AIM-1138

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 22 (28-IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form 2(pgs) ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages (by page number): _____

Photographs/Tonal Material (by page number): 3

Other (note description/page number):

Description :	Page Number:
IMAGE MAP (1) UNNUMBERED TITLE PAGE	
(2-22) PAGE #'ED 1-21	
(23) SCANCONTROL	
(24-26) TRGTS	
(27-28) DOD'S	

Scanning Agent Signoff:
Date Received: 1/19/95 Date Scanned: 1/26/95 Date Returned: 1/26/95

Scanning Agent Signature: Michael W. Cook

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AIM 1138	2. GOVT ACCESSION NO. AD-A215 375	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Stimulus familiarity determines recognition strategy for novel 3-D objects		5. TYPE OF REPORT & PERIOD COVERED memorandum 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Shimon Edelman, Heinrich Bulthoff, Daphna Weinshall		8. CONTRACT OR GRANT NUMBER(s) N00014-88-K-0164 DACA76-85-C-0010 N00014-85-K-0124
9. PERFORMING ORGANIZATION NAME AND ADDRESS Artificial Intelligence Laboratory 545 Technology Square Cambridge, MA 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, VA 22209		12. REPORT DATE July 1989 13. NUMBER OF PAGES 21
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Information Systems Arlington, VA 22217		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

psychophysics
vision
recognition
perceptual learning

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Everyday objects are more readily recognized when seen from certain representative, or canonical, viewpoints than from other, random, viewpoints. We investigated the canonical views phenomenon for novel 3D objects. In particular, we looked for the effects of object complexity and familiarity on the variation of response times and error rates over different views of the object. Our main findings indicate that the response times for different views become more uniform with practice, even though the subjects in our experiments received no feedback as to the correctness of their responses. In addition, the orderly dependency of the response time on the distance to a "good" view, characteristic of the canonical views. (Cont. on back)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 cont.

of a tradeoff between memory needed for storing specific-view representations of objects and time spent in recognising the objects.